Towards a Resource Grammar for Runyankore and Rukiga

David Bamutura
Chalmers Univ. of Tech. / Sweden
Mbarara Univ. of Sci. & Tech. / Uganda
bamutra@chalmers.se

Peter Ljunglöf
Univ. of Gothenburg / Sweden
Chalmers Univ. of Tech. / Sweden
peter.ljunglof@cse.gu.se

Abstract

In this paper, we present a formalisation of the grammar of two under-resourced Bantu languages: Runyankore and Rukiga (R&R). For this formalisation we use the Grammatical Framework (GF) and its Resource Grammar Library (GF-RGL) (Ranta, 2009; Ranta, 2011).

1 Background

Runyankore and Rukiga (R&R) are languages spoken in South-Western Uganda by about 6 million people (Simons and Fennig, 2018). They belong to the **JE10** zone (Maho, 2009) of the Niger-Congo Bantu language family. Previous work on R&R include: morphological analyzers by Katushemererwe and Hanneforth (2010a; 2010b), a Controlled Natural Language for Runyankore (Byamugisha et al., 2016) and a Noun pluralizer (Byamugisha et al., 2018). This work has been limited to small fragments of the languages but our approach aims at covering a significant amount of the lexicon and grammar of R&R which can be used as libraries for the development of richer NLP tools and applications.

Grammatical Framework (GF) is a grammar formalism based on type theory and a special-purpose functional programming language for defining grammars of both formal and natural languages (Ranta, 2004). GF is modular and highly expressive (Ljunglöf, 2004), making it appropriate for engineering libraries. It is well suited for working with under-resourced languages since it does not need any additional linguistic resources. The GF Resource Grammar Library (GF-RGL) is a set of of natural language grammars built with a common abstract syntax (Ranta, 2009). Resource grammars are important because they encourage division of labour between linguists who write libraries and domain experts who use them in applications (Cooper and Ranta, 2008). These grammars are domain-independent as shown by the different applications that have been built on top of them (Dymetman et al., 2000; Ranta et al., 2017; Lange, 2018).

2 Overview of the Morphology and Morphosyntax of Runyankore and Rukiga (R&R)

R&R are **mildly tonal**, **highly agglutinative** (e.g., the single word "tinkamureebagaho" (ti-n-ka-mureeb-a-ga-ho) is a sentence meaning "I have never seen him/her"), exhibit high instances of **phonological conditioning** and a **large Noun Class System** of 17–20 noun classes (Byamugisha et al., 2016) which is largely responsible for a complex concordial system of agreement among phrasal categories and larger syntactical categories combining them. This makes the languages complex to deal with and since they are under-resourced, a good approach is using rule-based/symbolic approaches, hence the choice of Grammatical Framework. We begin by describing R&R Morphology with special focus on Morphosyntax.

Nominal Morphology The morphological structure of nouns in R&R consists of two parts, a **class prefix** and a **noun stem**. The class prefix is further divided into an **Initial Vowel (IV)** and a **Noun Class Particle (NCP)** (Mpairwe and Kahangi, 2013b). The NCPs determine the noun class of the noun. The **noun stem** usually bears the bulk of the semantic meaning of the noun. Twenty noun classes for Runyankitara are suggested in (Katushemererwe and Hanneforth, 2010b). They use a numbered system

Universal Tense	Tense in R&R	Polarity	"To see"	Generalization
Present	Memorial Present	Positive	S-áá-reeb-a	S-TM-Rad-FV
		Negative	ti-S-áá-reeb-a	Pneg-S-TM-Rad-FV
	Experiential Present	Positive	S-∅-reeb-a	S-∅-Rad-FV
		Negative	ti-S-∅-reeb-a	Pneg-S-∅-Rad-Fv

Table 1: Table showing how different morphemes are combined to form a verb: Pneg = Primary Negative Marker, S = Subject Marker,TM = Tense Marker, ∅ = absence of TM, Rad = Radical and FV = Final Vowel.

of classification as opposed to the system of Noun Class particles used by Mpairwe and Kahangi (2013a; 2013b).

Verbal Morphology In Meeussen's (1967) original construction, the Bantu verbal unit consists of a **pre-stem** and a **stem**. The stem is further divided into a **base** and a **final vowel (FV)**. The base is also divided into a **radical (Rad)** and **extensions**. Further subdivisions in each of these parts results into 11 slots (Turyamwomwe, 2011), each with a set of morphemes that may appear in a particular slot for a particular purpose such as negative polarity (**Pneg / Sneg**), subject (**S**), object, tense, aspect and other markers. Regular verbs can be classified into four functional categories: Imperatives, Subjunctives, Perfectives and Infinitives. They can be rendered in active or passive voice and within each voice, the verb can take the form of Simple, Prepositional and Causative. In the verbal unit of R&R, Tense and Aspect (T/A) are marked using particular morphemes which may be simple or compound. Traditionally, tense is divided into Past, Present and Future. However, in R&R the past is split into Remote Past, Near Past and Immediate Past (Turyamwomwe, 2011). The present tense is divided into Universal Tense and Continuous / Progressive which are similar to Muzale's (1998) Experiential and Memorial Present. The Future is divided into Near and Far / Remote Future. As an example, Table 1 shows how different morphemes are combined to form a verb for the present tense.

Determiners and Adjectives In R&R it is impossible to express the definite and indefinite articles as distinct words. However, definiteness can be expressed morpho-syntactically using the Initial Vowel on the noun and other constituents in the noun phrase (Asiimwe, 2007). There are two types of adjectives: those that can stand on their own in a sentence or phrase and those that require a concord to be affixed to the stem. Among the latter, there are three types, adjectival stems whose concord is conjunctive with the stem and two others where the concord is disjunctive.

3 GF-RGL for Runyankore & Rukiga

We started with a basic resource grammar for our implementation and later extended it to a full resource grammar. We have implemented 22 syntactic categories and 46 grammatical functions which constitute the non-trivial aspects of the grammar. Nouns, Adjectives, Adverbs, Verbs and their phrasal counterparts have been implemented up to the level of Sentence, albeit in a selective manner. We have implemented morphological paradigms[1] for noun pluralization and partial verb conjugation based on the 38 rules in Mpairwe and Kahangi (2013a) for converting a verb in the imperative form into the perfective which forms the basis for predicting morphemes in the remaining slots after the radical. Despite the existence of 24 possible slots and the fact that those considered less important with reason may be ignored as explained in section 4 under discussion, full conjugation requires 4 GF-tenses where each can be either Simultaneous or Anterior (2), Positive and Negative Polarity (2), 35 Noun-Class particles for the Subject Marker tentatively bringing the number to 560 verb forms without putting into consideration possibilities of the verbal extensions after the radical. Attempting full verb conjugation leads to prohibitively large tables due to complexity of the verbal unit and hence makes GF fail to compile. However, by delaying the formation of the full surface form of the verb to the Sentence level, we were not only able to overcome creation of large tables, but also to avoid carrying them around from verbal lexical categories up through the various levels of phrasal categories i.e. verb phrases, declarative clauses,question clauses and relative

[1] A special function that computes all possible inflection forms with minimum information from the lexical item, usually the base form. It is well suited for regular forms of any lexical category but can be used with graceful degradation to the worst case where all forms are given but the grammar builder uses a single overloaded function.

clauses before reaching the Sentence (S) where the Tense, Anteriority[2] and Polarity are considered as parameters for its formation. We keep the different kinds of agreement concords in table structures that can be called upon when needed. One problem is that GF-RGL uses a universal 8 Tense/Aspect system, whereas R&R has in total 14 Tenses/Aspects. To solve this we devised a mapping between the two systems, and plan to implement the additional 6 as language extensions in the GF-RGL.

4 Discussion and Future Work

Discussion In order to maintain multilinguality, the abstract syntax of the GF-RGL restricts the grammatical aspects that can be implemented for all languages. For example, in our treatment of the verb, we ignore the use of the direct and indirect Object Markers because use of such markers would require anaphora resolution, which occurs at the discourse rather than the syntactic level hence the omission of the markers in Table 1 above. However, GF-RGL is flexible enough to allow the grammarian to implement language specific features as extensions, which we have done for structural words and intend to do for other syntactic categories. Phonological conditioning is a particular problem for highly agglutinative languages such as R&R which we have managed to solve in our smart paradigms[3] (Détrez and Ranta, 2012) for nouns. We plan to work on a global solution for other categories in future work, by producing a morphological analyzer and generator for R&R.

Our successful implementation of the basic resource grammar and the proper treatment of tense, aspect and polarity at the sentence level gives us confidence that implementation of the complete GF-RGLs will be possible with less problems. Since the grammars of R&R are very similar, as evidenced by their sharing of the same dictionaries (Taylor and Yusuf, 2009; Mpairwe and Kahangi, 2013a) and grammar books (Morris and Kirwan, 1972; Mpairwe and Kahangi, 2013b) and a high lexical similarity of 84%-94% (Turyamwomwe, 2011; Simons and Fennig, 2018) we can exploit GF's ability for making generalizations so that they can be reused.

Generally, GF-RGLs are useful in the development of multilingual applications such as localization of software applications, Multilingual Document Authoring (Dymetman et al., 2000), low-coverage multilingual translation (Ranta et al., 2010), domain specific dialogue systems such as music players (Perera and Ranta, 2007), Computer-Assisted Language Learning (CALL) (Lange, 2018; Lange and Ljunglöf, 2018a; Lange and Ljunglöf, 2018b) etc.

Despite the initial exposure to learning R&R in the first three years of primary school[4] and the existence of dictionaries, grammar books and an orthography, R&R largely remains a spoken, rather than written language even among those literate in English. Only a few study the language to a level sufficient to achieve proficiency in writing which implies lack of continuity in learning the grammar of the language. Our immediate motivation is therefore to utilise the GF-RGL for R&R to leverage the work done by Lange (2018) on CALL for the Latin language in order to build, localize and improve tools that can be used to create automatic exercises for learning R&R grammar to higher levels of profficeincy.

Future Work In the near future we plan to complete the RGL for the two languages and cater for the similarities, and to collaborate with other researchers working on Bantu languages in GF. In addition we will build application grammars to demonstrate the usefulness of the GF-RGLs developed, as well as treebanks and other linguistic resources for the two languages.

Acknowledgements

This work was supported by the Sida / BRIGHT Project 317 under the Makerere-Sweden Bilateral Research Programme 2015-2020. In addition, we would like to thank the organisers of the Widening NLP (WiNLP) workshop for offering the first author a travel grant worth USD 2344 to attend the workshop colocated with the Association for Compuational Linguitics (ACL) conference 2019.

[2] Whether tense has an auxiliary *have* or not

[3] A special paradigm requiring only the base form that works by leveraging both morphological and lexical knowledge.

[4] English becomes the official language of instruction and examination from the fourth year on, severely limiting the continued study of R&R to higher levels of proficiency.

References

Allen Asiimwe. 2007. *Definiteness and Specificity in Runyankore-Rukiga*. Ph.D. thesis, Stellenbosch University, South Africa.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2016. Bootstrapping a Runyankore CNL from an isiZulu CNL. In Brian Davis, Gordon J. Pace, and Adam Wyner, editors, *Controlled Natural Language*, pages 25–36. Springer International Publishing.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns across agglutinating Bantu languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2633–2643. Association for Computational Linguistics.

Robin Cooper and Aarne Ranta. 2008. Natural languages as collections of resources. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Relating Dialogue Coordination to Language Variation, Change and Evolution*. College Publications, London.

Grégoire Détrez and Aarne Ranta. 2012. Smart paradigms and the predictability and complexity of inflectional morphology. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–653, Avignon, France, April. Association for Computational Linguistics.

Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. Xml and multilingual document authoring: Convergent trends. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING 00, pages 243–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fridah Katushemererwe and Thomas Hanneforth. 2010a. Finite state methods in morphological analysis of Runyakitara verbs. *Nordic Journal of African Studies*, 19(1):1–22.

Fridah Katushemererwe and Thomas Hanneforth. 2010b. Fsm2 and the morphological analysis of Bantu nouns – first experiences from Runyakitara. *International Journal of Computing and ICT research*, 4(1):58–69.

Herbert Lange and Peter Ljunglöf. 2018a. MULLE: A Grammar-based Latin Language Learning Tool to Supplement the Classroom Setting. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA '18)*, pages 108–112, Melbourn. Australia. Association for Computational Linguistics.

Herbert Lange and Peter Ljunglöf. 2018b. Putting control into language learning. In *CNL 2018: Sixth International Workshop on Controlled Natural Language*, volume 304 of *Frontiers in Artificial Intelligence and Applications*, pages 61–70, Maynooth, Ireland. IOS Press.

Herbert Lange. 2018. *Computer-Assisted Language Learning with Grammars. A Case Study on Latin Learning*. Licenciate thesis, University of Gothenburg, Sweden.

Peter Ljunglöf. 2004. *Expressivity and complexity of the grammatical framework.* Ph.D. thesis, University of Gothenburg, Sweden.

Jouni Filip Maho. 2009. NUGL Online: The online version of the New Updated Guthrie List, a referential classification of Bantu languages. Technical report, University of Gothenburg. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.603.6490`.

Achille Emile Meeussen. 1967. Bantu grammatical reconstructions. *Africana Linguistica*, 3(1):79–121.

H. F. Morris and Brian Edmond Renshaw Kirwan. 1972. *A Runyankore grammar, by H. F. Morris and B. E. R. Kirwan*. East African Literature Bureau Nairobi, [rev. ed.] edition.

Y. Mpairwe and G.K. Kahangi. 2013a. *Runyankore-Rukiga Dictionary*. Fountain Publishers, Kampala.

Y. Mpairwe and G.K. Kahangi. 2013b. *Runyankore-Rukiga Grammar*. Fountain Publishers, Kampla.

Henry R T Muzale. 1998. *A Reconstruction of the Proto-Rutara Tense / Aspect System*. Ph.D. thesis, Memorial University of Newfoundland, Canada.

Nadine Perera and Aarne Ranta. 2007. Dialogue system localization with the gf resource grammar library. In *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*, SLP '07, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aarne Ranta, Krasimir Angelov, and Thomas Hallgren. 2010. Tools for multilingual grammar-based translation on the web. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 66–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aarne Ranta, Krasimir Angelov, Robert Höglind, Christer Axelsson, and Leif Sandsjö. 2017. A mobile language interpreter app for prehospital/emergency care. In *Medicinteknikdagarna 2017*, Västerås, Sweden.

Aarne Ranta. 2004. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.

Aarne Ranta. 2009. The GF Resource Grammar Library. *Linguistic Issues in Language Technology*, 2(1).

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Gary F. Simons and D. Fennig, Charles. 2018. *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, Twenty-first edition. Online version:`http://www.ethnologue.com`.

Charles V Taylor and Mpairwe Yusuf. 2009. *A simplified Runynakore-Rukiga-English DICTIONARY*. Fountain Publishers, Kampala, revised ed edition.

Justus Turyamwomwe. 2011. Tense and aspect in Runyankore-Rukiga, linguistic resources and analysis. Master's thesis, NTNU – Norwegian University of Science and Technology.